Improved Text mining for bulk data using Deep learning approach

Indumathi A
PG Scholar,
Department of Computer Science and Engineering,
Sri Ramakrishna Engineering College,
Coimbatore.

Perumal P
Professor,
Department of Computer Science and Engineering,
Sri Ramakrishna Engineering College,
Coimbatore.

Abstract- Text document clustering and similarity detection is the major part of document management, where every document should be identified by its key terms and domain knowledge. Based on the similarity, the documents are grouped into clusters. For document similarity calculation there are several approaches were proposed in the existing system. But the existing system is either term based or pattern based. And those systems suffered from several problems. To make a revolution in this challenging environment, the proposed system presents an innovative model for document similarity by applying back propagation time stamp algorithm. It discovers patterns in text documents as higher level features and creates a network for fast grouping. It also detects the most appropriate patterns based on its weight and BPTT performs the document similarity measures. Using this approach, the document can be categorized easily. In order to perform the above, a new approach is used. This helps to reduce the training process problems. The above framework is named as BPTT. The BPTT has implemented and evaluated using dot net platform with different set of datasets.

1. INTRODUCTION

The capacity of storage data becomes huge amount of the technology of computer hardware develops. So amount of data is increasing exponentially, the information required by the users become varies. Actually users deal with textual data more than the numerical data. It is very difficult to apply techniques of data mining to textual data instead of numerical data. Text miming [1] is finding interesting regularities in large Textual datasets. The text mining studies are gaining more importance recently because of the availability of the increasing number of the documents from a variety of sources. Which include unstructured and semi structured information. The main functions [2] of the text mining include text summarization, text categorization and text clustering. The Text of this paper is restricted to text categorization.

“Text mining” is increasingly being used to denote all the tasks that, by analyzing large quantities of text and detecting usage patterns, try to extract probably useful (although only probably correct) information.

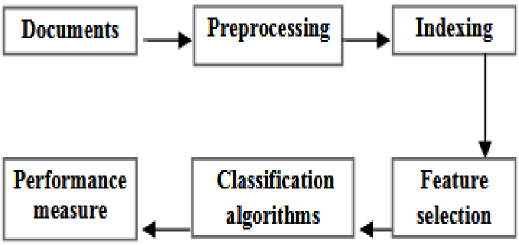

Fig.1.1 Document classification process

Deep learning approach [3] are representation learning methods with multiple levels of representation, but nonlinear modules that methods transforms the representation at one level (starting with the raw input) into a higher representation slightly more abstract level, with the composition of enough such transformations, and very complex functions can be learned. Deep learning approach of learning algorithm, feature extraction can improve the accuracy of learning algorithm and shorten the time. Selection from the document each part can reflect the information on the text classification, and the calculation of weight is called the text feature extraction.

2. RELATED WORK

In the recent years, the progress of web and social network technologies have led to a massive interest in the classification of text documents containing links or other meta-information and many studies on classification algorithms have been done by many researches. In this section we will do a review to these works and show the focus points of them. As we will see, the novelty of our work is appears by studying almost all the modification and improvements to each algorithm. Focused [4] on specific changes which are applicable for the text classification. They used, as text classification algorithms, Decision Trees, Pattern (Rule) based Classifiers, SVM Classifiers, Neural Network Classifiers, Bayesian (Generative) Classifiers, nearest neighbor classifiers, and genetic algorithm based classifier. They are discussed the methods used for in text classification and described these methods for text classification. To text classification [5] process of text classification as well as the classifiers and tried to compare the some existing classifier on basis of few criteria like time complexity, principal and performance. The theory and methods of text classification and text mining, the important algorithms that are text classification. In features [6] of each category by using the information. In this performance for this algorithm was reasonable where they showed that feature selection in the decision tree algorithm was particle effective in dealing with the large feature sets common in text categorization. They used the feature extraction and modified the used algorithm. They are many improvements to the well-known algorithms for text classification. The improvements in algorithm can be modification/addition to the algorithm and the learner.

3. PROPOSED SYSTEM

In this proposed method derives text similarity from semantic and syntactic information contained in the similarities text. A text is considered to be a sequence of words each of which carries useful information. The words along with their combination structure make a text convey a unique meaning.

Clustering is the most widely used technique in text mining process. It organizes a large quantity of disordered text documents into a small number of meaningful and sticking together clusters, they provides the foundation for something for intuitive and informative navigation and browsing mechanisms. Text-clustering is to divide a collection of text- documents into several categories so that documents in the same concept describe that identical topic such as classical music. Text Clustering efficiently groups documents with similar collection into same cluster. Similarity between objects is measured within the use of similarity function.

The back propagation based Time algorithm is used for fast document similarity analysis. In a recurrent neural network, errors can be propagated further, i.e. more than 2 layers, in order to capture longer history information. This process is usually called unfolding. The recurrent weight in an unfolded RNN is duplicated spatially for an arbitrary number of time steps, here referred to as τ . In accordance with Equation 1, errors are thus propagated backward as:

$$\delta_{pj}(t-1)=\sum_{h}^{m}\delta_{ph}(t)u_{hj}f'(s_{pj}(t-1))$$

Where,
h is the index of hidden node at time t. The ignorance deltas of higher layer weights are calculated recursively. After obtaining all the error deltas, weights are folded back adding up to one big change for each unfolded weights.
Figure 3.1 shows the classification using similarity. The proposed algorithm consists of two stages; the first stage is clustering, and the second stage is flow level classification.

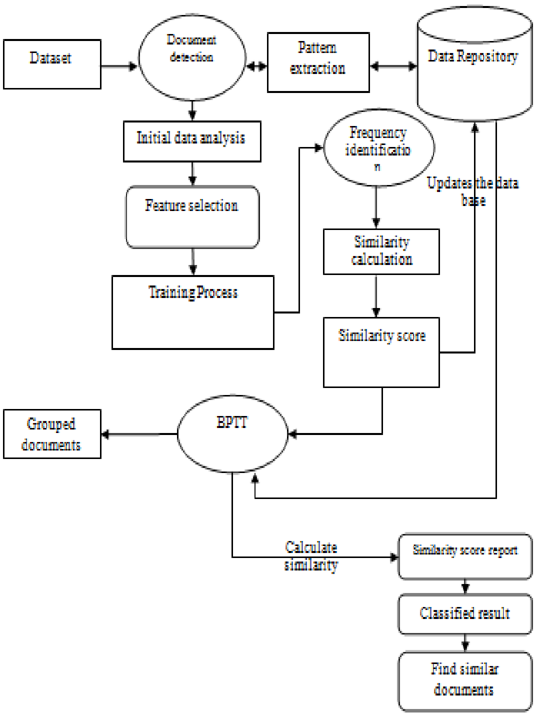

Figure 3.1 Flow of BPTT

The back propagation training algorithm similar documents from the big data environment. The mathematical method used to calculate derivatives of chain rule. This is a training algorithm for updating network weights to decrease error.

The BPTT has the following steps.

- The pattern of input and propagate it through time to get on output
- Analyze the predicted outputs to the expected outputs and calculate the error.
- Calculate the derivative weights of the error.
- Adjust the weights to minimize the error.
- Repeat.

4. DATA SET

The proposed system used real-time and synthetic datasets. Different corpus adopts different rules and models. Some have documents with specialized vocabulary containing words that are repeated frequently. On the other hand, corpus derived from certain sources exhibit creative writing style with word occurrences seldom repeated in their documents. Further details, including discussion of previous versions of the collection (e.g. Reuters-22173), are available in the website. The dataset is available http://www.research.att.com/~lewis/reuters21578.html and ftp:://canberra.cs.umass.edu/pub/reuters. It has 90 specialized categories. All the 90 categories can be used in the experiments.

5. RESULT AND DISCUSSION

Assessment of overall performance: In this subsection, the report gives the results and overall performance of the proposed BPTT model. So, the first process is comparing its accuracy with that obtained by BPTT. Then the next illustrate the variety of prior solutions present in the final iteration. Finally this gives the salient performance parameters of the best pattern obtained for each dataset and compare them with previously reported results.

Comparison with BPTT Model: In order to compare with the BPTT approach with a existing system, this chapter conducted BPTT based document grouping process using patterns of each document in each of the data sets. The existing BPTT was trained and tested for each corpus. Table 5.1 tabulates the accuracy results obtained for the two approaches.

a) The proposed collaborative approach performs comparatively better than BPTT for both datasets of each corpus. The average accuracy for the collaborative method is 95.55% as compared with 81.44% with the BPTT method, thus giving an improvement of 25%.

Table 5.1 Performance Comparison between existing and BPTT approaches.

Datasets	**Accuracy using existing system (%)**	**Accuracy using BPTT (%)**
R21578	86	96.5
Dataset1	84	97
Dataset2	83	96

b) In cases where the BPTT method gave acceptable results, i.e. 86% for the R21578 dataset and 84.5 % for the Dataset1, the approach enhanced it in both cases to 96.5% and 97% respectively.

c) For the synthetic and large dataset 2, the BPTT approach led to rather poor results which were dramatically improved with a collaborative approach. For instance, the classification accuracy of the Dataset2 was only 83% using existing approach. This improved to as much as 96% with the BPTT approach. This is because the DC system was able to utilize the context based pattern maximally in the domain corpus.

Nowadays, document classification in system requires high detection rate and low false alarm rate, thus the research compares accuracy, detection rate and false alarm rate, and lists the comparison results of various documents.

Table: 5.2 Performance comparison table.

Metrics	**Existing**	**Proposed**
Similarity calculation Time(ms)	4.3	2.2
Efficiency	Ordinary	Better
Accuracy (%)	90.7	97.5

The comparison between existing and proposed system based on the Training time. The training time of the other classification algorithms with the proposed system.

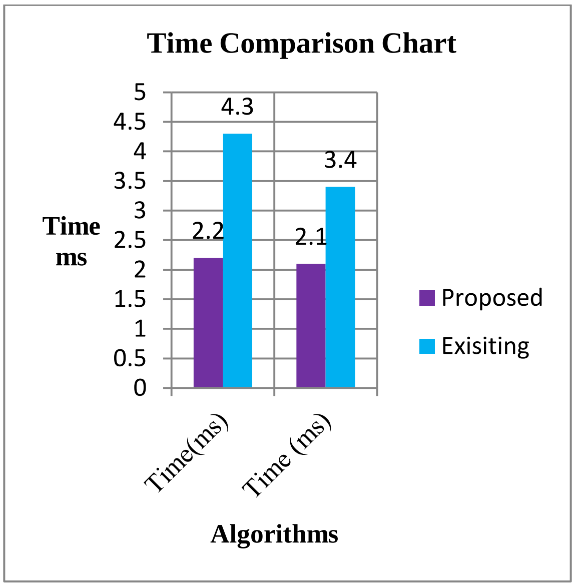

Fig: 5.2 Time comparison between existing cosine similarity and proposed BPTT

6. CONCLUSION

Mining is a significant research area which is gaining an increasing popularity in the recent years. The similarity between the text documents is an important operation of text mining. Text Classification is an important application area in information retrieval, text mining. Because classifying millions of text document manually is an expensive and time consuming task. In order to reduce the training process, a BPTT approach is implemented in this project. The system proposed an effective method with various patterns for document grouping. This paper also performed the similarity measure for the given two documents based on its external and gathered features.

7. REFERENCES

[1] Hung Chim and Xiaotie Deng, (2008) “Efficient Phrase-Based Document Similarity for Clustering,” IEEE Transactions on Knowledge and Data Engineering, Vol. 20, Issue. 9, pp. 1217 – 1229.

[2] Wael H. Gomaa Aly A. Fahmy,(2013) “A Survey of Text Similarity Approaches,” International Journal of Computer Applications, Vol.68, pp.1-13.

[3] B.Pangand L.Lee, (2008) “Opinion mining and text analysis,” International Conference on Information Technology, Vol.2, Issue.2, pp.1–35.

[4] Pablo Basanta-Val, Neil C.Audsley, Andy J. Wellings, Ian Gray, and Norberto Fernandez-Garcıa, (2016) “Architecting Time-Critical Big-Data Systems,” IEEE Transactions on Big Data, Vol. 2, pp.1- 4.

[5] Amita Verma, Ashwani kumar, (2014) “Performance Enhancement of K-Means Clustering Algorithms for High Dimensional Data sets,” International Journal of Advanced Research, Vol. 4, Issue. 1, pp 791-796.

[6] Y.Lu,C.Zhai, and N.Sundaresan, (2009) “Rated aspect summarization Of short comments,” International Conference on World Wide Web, Vol.1, pp.131–140.

[7] Potts C, (2010) “From frequency to meaning: vector space models of semantics,” Journal Artif Intell, Vol.4, Issue.3 ,pp.1-8.

[8] K.Fanand, C.H.Chang,(2010) “Text-oriented contextual advertising,” Knowledge and Information Systems, Vol.23, Issue.3, pp. 321–344.

[9] Manning CD, Raghavan P, Schutze H, (2008) “Introduction to information retrieval,” IEEE Conference on Information Technology, Vol.6, Issues.2, pp. 279–288.

[10] Wellings AJ, Audsley NC, Basanta-Val P, Fernndez Garca N, (2015) “Improving the predictability of distributed stream processors,” Science Direct on Computer Application, Vol.52, pp. 22–36.

[11] M.Hu and B.Liu, (2004) “Mining and summarizing customer reviews,” in KDD2004, pp.168–177.

[12] Kumar S, Toshniwal D (2016), “A novel framework to analyze road accident time series data,” Journal of Big Data, Vol.3, pp.1-8.

[13] H. Becker, M. Naaman, and L. Gravano,(2010) “Learning similarity metrics for event identification in social media,” The third ACM international conference on Web search and data mining, Macau, China, pp.131-142

.

[14] Kumar S, Toshniwal D, (2016) “Analysis of hourly road accident counts using hierarchical clustering and cophenetic correlation coefficient,” Journal of Big Data, Vol.3(1), pp.1–11.

[15] Michie MG, (1982) “Use of the bray-curtis similarity measure in cluster analysis of foraminiferal data,” Journal of Big Data, Vol.14, pp.661–667.